Robust and scale-free effect sizes for non-Normal two-sample comparisons, with applications in e-commerce

David A. Wooff[a*] and A. Jamalzadeh[b]

[a] *Durham University, Department of Mathematical Sciences, Stockton Road, Durham DH1 3LE, UK*
[b] *Summit Media Ltd, Albion Mills, Albion Lane, Willerby, Hull, HU10 6DN, UK*

May 31, 2013

Abstract

The *effect size* (ES) has been mainly introduced and investigated for changes in location under an assumption of Normality for the underlying population. However, there are many circumstances where populations are non-Normal, or depend on scale and shape and not just a location parameter. Our motivating application from e-commerce requires an ES which is appropriate for long-tailed distributions. We review some common ES measures. We then introduce two novel alternative ES for two-sample comparisons, one scale-free and one on the original scale of measurement, and analyse some theoretical properties. We examine these ES for two-sample comparison studies under an assumption of Normality and investigate what happens when both location and scale parameters differ. We explore ES for phenomena for non-Normal situations, using the Weibull family for illustration. Finally, for an application, we assess differences in customer behaviour when browsing E-commerce websites.

Keywords: effect size; two-sample comparison; non-Normal distribution; Weibull; quantile function; e-commerce; long-tail distribution.

1 Introduction

Classical hypothesis-testing is the standard way of using experimental data to test existence of a phenomenon Gigerenzer [1993], Ledesma et al. [2009]. However, statistical significance does not provide information about the magnitude, or *effect size* (ES), of the phenomenon, whereas this is usually the focus of research Krueger [2001]. Provision of such measures is thus a necessary complement to hypothesis-testing Thompson [1998]. Indeed, each of the quantities: p-value, sample size, a measure of ES, and test power, is typically a function of the other three Cohen [1992], Descôteaux [2007] and so in quantifying and interpreting experimental research we should consider them together.

The p-value is not a direct measure of ES. One reason is that statistically significant findings are not always practically significant, because taking a large enough sample size will result in a small p-value. Conversely, a statistical test with weak power, perhaps because of small sample size, may not appear as statistically significant, but the measured effect relative to background variation may be deemed to be of practical significance Wilson Van Voorhis and Morgan [2007]. For this reason, appropriate specification of both statistical significance *and* the magnitude of effect, is required to provide inference of practical utility Thompson [1998].

Measures of ES have been available for decades, but mainly limited to meta-analysis for combining estimates from different studies Keselman et al. [1998], perhaps because they have been largely developed for simplistic situations, such as changes in location under assumptions of Normality. However, there are many circumstances where populations are non-Normal, or depend on scale and shape and not just a location parameter. There are non-parametric measures of ES which require weaker assumptions, but often as *dominance* measures. One distribution F *dominates* a second distribution G when each quantile of F is

*Corresponding author. Email: d.a.wooff@durham.ac.uk

larger than the corresponding quantile of G. However, there are many settings in which two distributions do not necessarily dominate each other, but exhibit contrasting differences for different parts of the range. For example, when comparing two lifetime distributions for a survival analysis, the survival function for one group may be initially larger than that for the other group, but then cross over and become smaller. Classical ES do not truly describe these differences.

In Section 2, we briefly review existing ES for two-sample comparison studies. In Section 3 we motivate and introduce two robust ES measures based on quantile differences and use these to examine changes in location under a Normality assumption, and thence to changes in location and scale. In Section 4 we show how the proposed ES behaves for general comparisons of two distributions. In Section 5 we explore ES under Weibull distributions to illustrate comparisons for highly non-Normal populations. For a practical application, we compute ES in the context of identifying patterns in web browsing behaviour in Section 6. In Section 7 we show how a bootstrap approach may be used to derive confidence intervals for our proposed ES measures. Algorithms to construct the ES measures we propose are available in the appendix.

2 Effect sizes for the two-sample comparison

2.1 Cohen's d and d_r

The most familiar ES is Cohen's d Cohen [1977], estimated by $d = (\bar{x}_1 - \bar{x}_2)/s$, where $\bar{x}_1$, $\bar{x}_2$ are the sample means, and s is an estimate of the common standard deviation. This expresses the intuitively appealing concept that the magnitude of effect is the difference between the centres of two populations relative to a measure of individual variation. There is often implicitly the assumption that the underlying populations are Normal.

Cohen offered practical rules to interpret d: an ES of 0.2 to 0.3 standard deviations is deemed a *small* effect; ES ≈ 0.5 is a *medium* effect; and ES ≈ 0.8 is a *large* effect. He warned that such criteria are relative and interpretations must take into account the content, purpose, and method of research, except perhaps in the context of research with entirely novel variables Lenth [2001]. The value of this rule to applied research has been questioned, since the practical importance of ES also depends on other quantities, such as the effectiveness of alternative treatments and cost-benefit analysis of the treatment Hodges and Olkin [1985].

Cohen Cohen [1977] also proposed a correlation form, d_r, of his d index, where d_r is the correlation between numerical variable Y and binary variable X:

$$-1 < d_r = d/\sqrt{d^2 + (1/pq)} < 1, \tag{1}$$

where p and q are the proportions of subjects belonging to the two groups indicated by X. Cohen's d_r is large when d is large and also when p and q are similar. d_r is bounded, which may facilitate interpretation. For a critical review of the ingredients to d as an ES, and a proposed alternative d_2, see Cahan and Gamliel [2011].

2.2 Simple robust alternatives

There have been many attempts to extend Cohen's d to more complicated situations, for example weakening the assumption of common variance. One recommendation Glass et al. [1981], which scales $(\bar{x}_1 - \bar{x}_2)$ by the estimated standard deviation of the control group, simply ignores that we should prefer an ES which explicitly incorporates heteroscedasticity. Commonly used ES are not robust. Small changes in the tails can substantially inflate variance estimates and thus decrease Cohen's d. A well-known example is based on the contaminated Normal distribution Tukey [1960], Wilcox [2005], Wilcox and Tian [2011]. Suppose that $H(x) = 0.9\Phi(x) + 0.1\Phi(x/10)$, where Φ is the standard Normal cdf. Now consider two groups, both Normally distributed and with common variance $\sigma^2 = 1$, and suppose their means are $\mu_1 = 0$, $\mu_2 = 0.8$, so that $d = 0.8$, a *large* ES. For the contaminated Normals with the same means, Cohen's d falls to $d = 0.59$, a *medium* ES.

Means and variances may be replaced by robust alternatives, for example using 20% trimmed means and a Winsorized variance Algina et al. [2005]. There is a loss of tail information when trimming, so this may be a serious drawback when comparison in tails is important, as for the example we show in section 6.

See Hedges and Friedman [1993] for an analogue for Cohen's d when the focus is on comparing the tails or some other feature of a distribution rather than the centre.

2.3 Other alternatives

The literature contains many other proposals, with a range of advantages and disadvantages. The Common Language ES can be calculated for different probability distributions and under different assumptions and has the advantage of representing ES using a common probability scale Anderson and Berry [2009], Ledesma et al. [2009]. It is often employed as a measure of numerical dominance in location to judge whether one distribution is generally larger or smaller than another. The *probability of superiority* ES Grissom and Kim [2005] is similar, but based on sample ranks. This measure assumes similarity of shape in the underlying populations, but is not robust to heteroscedasticity Mann and Whitney [1947], Grissom [1994].

Non-overlap ES can be obtained by comparing the percentiles of populations. Three such measures were proposed by Cohen Cohen [1977] under an assumption of Normality with equal variation.

Cliff Cliff [1993] introduced a simple non-parametric ES, δ, which is computed by counting the number of occurrences of an observation from one group having a higher response value than an observation from the second group, and vice-versa. $\delta = \pm 1$ indicates full separation between the groups, whereas $\delta = 0$ indicates full overlap. When the data do not follow the Normal distribution or where the variable under study corresponds to an ordinal level of measurement, Cliff's δ has been preferred to Cohen's d or d_r Hess and Kromrey [2004]. As a non-parametric measure, its interpretation is unaffected when the assumptions of Normality or homoscedasticity are violated Coe [2002]. However, δ is perhaps more suitable as a crude dominance measure.

Wilcox & Tian Wilcox and Tian [2011] proposed an *explanatory power* ES based on special case measures that reflect the proportion of variance in a response variable Y accounted for by an explanatory variable X using regression. One advantage is that it can be generalized to multi-sample problems. See Doksum and Samarov [1995] for a simple linear regression example, and Kulinskaya and Staudte [2006] for a robust analogue which however depends on sample size.

2.4 Graphical representations

Quantile and percentile plots have long been exploratory graphical tools for comparison studies Wilk and Gnanadesikan [1968], including suggestions Doksum and Sievers [1976], Doksum [1977] to use a graphical representation of the shift function for comparing groups. This involves making comparisons at a number of distribution features, rather than only location or scale. The magnitude of difference between two populations may also be explored using the ordinal dominance curve Darlington [1973], where the area under or above the curve may be used as a measure of dominance. For a better separation of points when observations are close to the $y = x$ line, the Tukey sum-difference graph has been recommended instead of the quantile plot Cleveland [1994]. Such plots provide a complete representation of the differences between two populations, but not a single summary ES.

3 Developing effect sizes for non-Normal data

There are many circumstances where populations are known or suspected to be non-Normal, and perhaps depend on scale and shape as well as location. Basing an ES on medians alone does not help: the medians of two distributions might be equal, but their tails might be very different Fleming et al. [1980]. Indeed, where two populations may be Non-Normal, or heteroscedastic, or nonhomomerous in other ways, traditional measures of ES can be misleading and do not provide sufficient information about the magnitude of the effect Grissom and Kim [2005], Wilcox [2005].

Our need is to construct an ES over the full distribution. Extending from the notion of the graphical comparison described above, one possibility for the two-sample comparison is to compare the *quantile functions* and *vertical quantile comparison functions* of two distributions F and G. In principle, these should provide general ES which are applicable to both non-Normal and Normal settings. We begin with some definitions.

For a probability distribution function F, the *quantile function* for F is given by

$$Q(p) = \inf\{x \in R : p \leq F(x)\}, \quad 0 \leq p \leq 1. \tag{2}$$

For discrete probability distribution functions, the quantile function returns the minimum value of x for which the statement holds. For random variables for which the cumulative distribution function (cdf) is continuous and strictly monotonic, $F\colon R \to (0,1)$, $Q(p)$ is the inverse of the cdf, and it is common to use F^{-1} as notation.

Suppose that F and G are continuous probability distribution functions. The *vertical quantile comparison function* for G with baseline F is

$$V_F^G(p) = G(F^{-1}(p)), \quad 0 \leq p \leq 1. \tag{3}$$

This may be used to represent the distance between two probability distributions Li et al. [1996]. If $F = G$ then $V_F^F(p)$ is the probability distribution function of the Uniform distribution. Thus, differences between $V_F^G(p)$ and the Uniform cdf equate to differences between F and G. One may plot the vertical quantile comparison function versus the probability to display the difference between two distributions Holmgren [1995].

The *vertical shifted function* for G with baseline F is defined to be $V_F^G(p) - p$. This summarizes differences between F and G at each point p because the Uniform cdf is $F(p) = p$.

3.1 Quantile absolute deviation

In this section, we extend the idea of comparing two distributions by their quantiles for the entire range of probabilities over $[0, 1]$. For two populations with cumulative distribution functions F and G, we define the *quantile absolute deviation* (QAD) as:

$$\mathrm{QAD}(F\ ,\ G) = \int_0^1 \left|\ F^{-1}(p) - G^{-1}(p)\ \right| \, dp, \tag{4}$$

where F^{-1} and G^{-1} are quantile functions for the two distributions. The QAD is a symmetric positive measure and may be interpreted as the average distance between the quantiles of the distributions. When two distributions are similar, their quantiles must also be similar and $\mathrm{QAD}(F\ ,\ G)$ will be small. The QAD satisfies the three divergence properties of a criterion:

1. Self similarity: $\mathrm{QAD}(F\ ,\ F) = 0$.
2. Self identification: $\mathrm{QAD}(F\ ,\ G) = 0$ if and only if $F = G$.
3. Positivity: $\mathrm{QAD}(F\ ,\ G) \geq 0$ for all F, G.

The space of quantile functions with the above mentioned distance is a metric space because the distance fulfills all the following axioms of a metric. For probability distribution functions F, G, and H:

1. Self identification: $\mathrm{QAD}(F\ ,\ G) = 0$ if and only if $F = G$.
2. Symmetry: $\mathrm{QAD}(F\ ,\ G) = \mathrm{QAD}(G\ ,\ F)$.
3. Triangle inequality: $\mathrm{QAD}(F\ ,\ G) + \mathrm{QAD}(G\ ,\ H) \geq \mathrm{QAD}(F\ ,\ H)$.

The QAD is not a scale-free measure as it has the same unit of measurement as the variable under investigation. Thus, $\mathrm{QAD}(F\ ,\ G) = 1$ implies that on average the quantiles of the variable with distribution F differ by one unit from the quantiles of distribution G. It may or may not be true that one of the distributions dominates the other, but if this is the case, the QAD may be interpreted directionally.

One may exclude the observations which lie in the far tails of the distribution, in order to eliminate the effect of outliers or extreme values on the ES. A $100\alpha\%$ trimmed QAD can be calculated using only observations which are placed between the $(\alpha/2)$th and $(1-\alpha/2)$th percentiles of the data, computed as:

$$K_\alpha(F\ ,\ G) = \frac{1}{(1-\alpha)} \int_{\alpha/2}^{1-\alpha/2} \left|F^{-1}(p) - G^{-1}(p)\ \right| \, dp. \tag{5}$$

This is more robust to outliers, but does not give a metric measure over the space of quantile functions and loses tail information.

3.2 Quantile comparison effect size

In this section, we introduce the notion of using a distance measure – statistical divergence – as an ES. Statistical divergence is a weaker notion than that of distance as it does not need to be symmetric. That is, the divergence from F to G is not necessarily equal to the divergence from G to F. The vertical shifted function $V_F^G(p) - p$ provides a basis for such a divergence. For two cumulative distribution functions F and G, define

$$Div(F \,\|\, G) = 2 \times \int_0^1 \Big| G(F^{-1}(p)) - G(G^{-1}(p)) \Big| dp = 2 \times \int_0^1 \Big| \, V_F^G(p) - p \, \Big| dp, \tag{6}$$

which takes values over the interval $[0, 1]$. Although this measure satisfies divergence properties, it is not necessarily symmetric. However, the corresponding average is, and this is what we propose to use as an ES. Thus, we define the *quantile comparison effect size* (QCES) as

$$\begin{aligned} \text{QCES}(F\,,\,G) &= \frac{1}{2} Div(F \,\|\, G) + \frac{1}{2} Div(G \,\|\, F) \\ &= \int_0^1 \Big| \, G(F^{-1}(p)) - p \, \Big| + \Big| \, F(G^{-1}(p)) - p \, \Big| \, dp. \end{aligned} \tag{7}$$

This is a bounded measure, giving values between 0 and 1, which facilitates its interpretation as an ES, unlike alternative unbounded divergence measures.

Standard divergence measures include Kullback-Leibler (KL), the relative entropy between two continuous probability density functions $f(x)$ and $g(x)$ Kullback [1968]:

$$D_{KL}(D \,\|\, G) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx. \tag{8}$$

This is commonly used as a measure of similarity between two density distributions. In a Bayesian context, this divergence and its variants may be used to measure the difference between prior and posterior distributions, and symmetrized versions may be obtained.

ES based on quantile comparison can be computed directly where F and G are known. Otherwise we may substitute the cdf by the empirical cdf. Thus, suppose we draw samples from distribution F and distribution G and estimate F by $\widetilde{F}$ and G by $\widetilde{G}$, the corresponding empirical cdfs. From these we may determine the empirical quantile functions $\widetilde{F}^{-1}$ and $\widetilde{G}^{-1}$ as appropriate.

3.3 Interpreting the Quantile Comparison effect size

Cohen's practical rules to interpret ES can be mapped into the QCES by evaluating the QCES for Normal distribution comparisons where we fix the scale at $\sigma_X = \sigma_Y = 1$ and consider location changes of sizes deemed by Cohen to be *small* ($\mu_X - \mu_Y = 0.2$), *medium* ($\mu_X - \mu_Y = 0.5$), and *large* ($\mu_X - \mu_Y = 0.8$). A graph showing the QCES for general choices for $d = \mu_X - \mu_Y$ is shown in Figure 1.

If we adopt similar thresholds, this suggests that QCES values of around 0.1 to 0.2 correspond to *small* effects; values around 0.2 to 0.4 correspond to *medium* effects; and larger values suggest *large* effects. A QCES of at least 0.7 represents the situation where two Normal distributions with the same scale mostly do not overlap. For more complicated comparisons with non-Normal distributions and possible changes in the values of several parameters, distributions may still contain substantial overlap with respect to quantiles. Having assessed the QCES over several plausible comparisons, we thus suggest the following guidelines. A value of QCES in (0.10,0.20) suggests a *small* effect; (0.20,0.30) suggests a *medium* effect; and 0.30 and above corresponds to a *large* effect. As Cohen suggests, such ES need to be interpreted in context; computation of a summary ES cannot replace detailed comparison of the two distributions. For the robustness illustration of §2.2 using contaminated Normal distributions, the corresponding values of the QCES are 0.26 and 0.21, indicating *medium* effect sizes on our scale in both cases.

3.4 Computation for some theoretical cases

The computation of QCES and QAD involves integration over a function of two quantile functions. As such, they rarely provide tractable analytic forms, particularly for distributions for which there is no simple

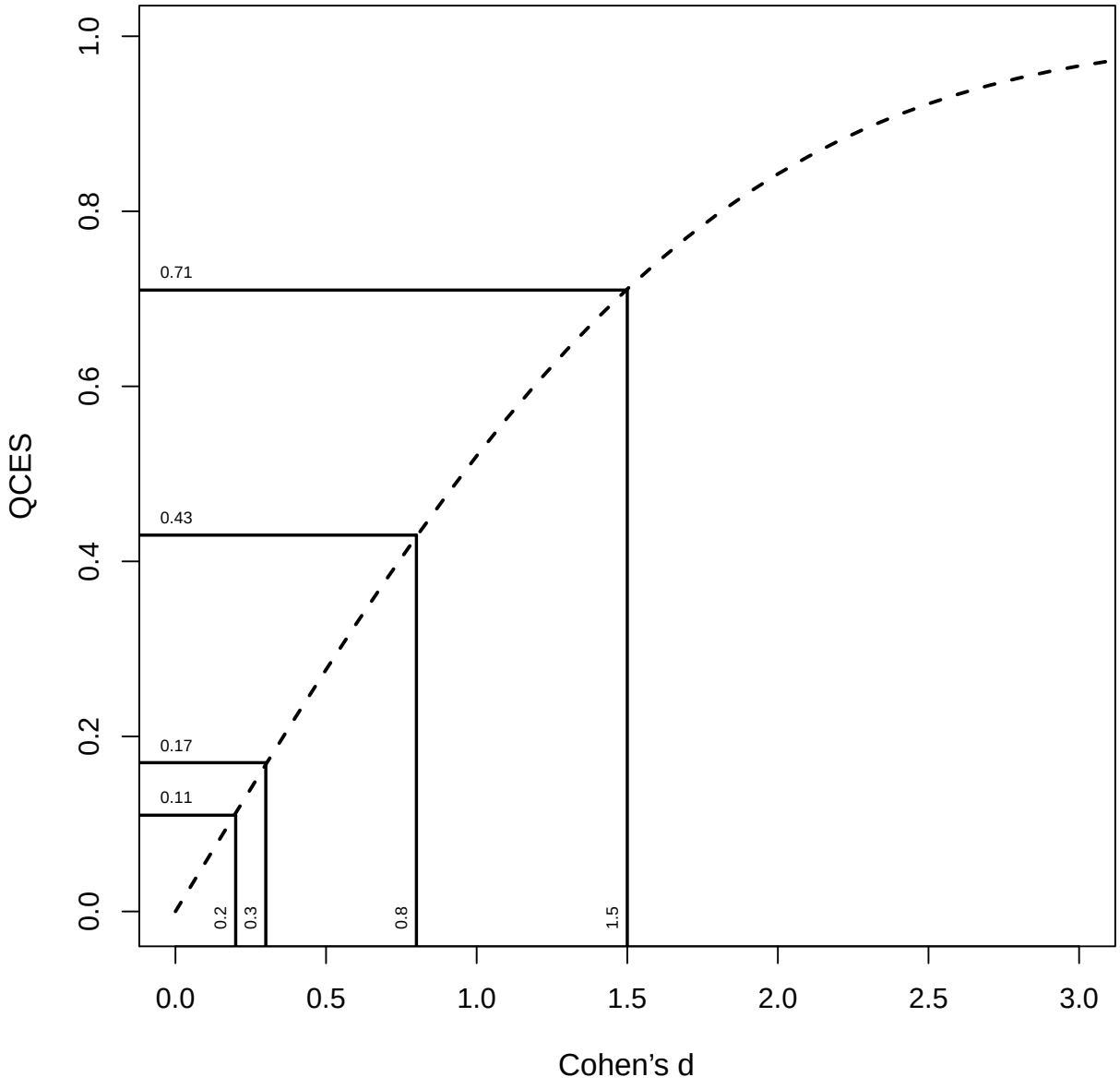

Figure 1: Values of the QCES corresponding to standard thresholds for Cohen's d ES.

expression for the quantile function; see Gilchrist [2000] for details of quantile functions for a large number of known distributions. However, it is informative to investigate the behaviour of these ES in terms of the parameters of distributions in some simple settings.

Uniform distribution with different locations: Suppose that $X \sim U(0\,,\,\alpha)$ and $Y \sim U(0\,,\,\beta)$ where $\beta \geq \alpha$. The quantile function for $X \sim U(a\,,\,b)$ is $F_X^{-1}(p) = (b-a)\,p + a$. So, using (2) the QAD is given by

$$\mathrm{QAD}(X\,,\,Y) = (\beta - \alpha)/2, \tag{9}$$

In this case, the QAD is simply the difference between the means of the two distributions. The QCES (7) is computed using $Div(X \,\|\, Y)$ and $Div(Y \,\|\, X)$, remembering the restriction to $\beta \geq \alpha$, giving

$$\mathrm{QCES}(X\,,\,Y) = 1 - \alpha/\beta. \tag{10}$$

For β close to α the QCES tends to zero, and for $\beta \gg \alpha$ the QCES approaches the upper limit, one.

In practice, we typically won't know the values of parameters and so must estimate them. One way to explore the behaviour of these ES is to treat these parameters as random variables, in which case these ES are random variables. Suppose, for example, that the parameters are independently $\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$ and $\beta \sim N(\mu_\beta, \sigma_\beta^2)$. In the above Uniform case, the QAD is Normal:

$$\mathrm{QAD}(X\,,\,Y) \sim N\Big(\frac{\mu_\beta - \mu_\alpha}{2}\,,\,\frac{\sigma_\beta^2 + \sigma_\alpha^2}{4}\Big) \tag{11}$$

suggesting that we would expect the ES to be the underlying difference between the two Uniform means, but with variation stemming from uncertainty about those means. The QCES is a simple linear transformation of a ratio of two independent non-zero mean Normal distributions, see Hinkley [1969]. In some cases this ratio distribution can be well approximated by another Normal distribution which can serve as the basis for exploration of behaviour: see Jamalzadeh [2010] for details.

Exponential distribution with different rates: Suppose that $X \sim Exp(\alpha)$ and $Y \sim Exp(\beta)$ where $\beta \geq \alpha$. The quantile function for $X \sim Exp(\theta)$ is $F_X^{-1}(p) = -\frac{1}{\theta}\,\ln(1-p)$. Thus, the QAD (4) and QCES (7) are:

$$\mathrm{QAD}(X\,,\,Y) = 1/\alpha - 1/\beta, \quad \mathrm{QCES}(X\,,\,Y) = (\beta - \alpha)/(\alpha + \beta),$$

so that the the QAD and the QCES relate to the difference and scaled difference in means, as is so for the Uniform case. As above, we could explore these ES as random variables, assuming Normal distributions for the α and β parameters.

4 Two-sample Normal distribution comparisons: simulation study

We now explore the QAD and the QCES for a number of situations. First, we assume that the underlying distributions are known and we examine ES for the Normal distribution comparison and, later, for the Weibull distribution comparison, varying parameter choices in each case. In a further section we examine ES calculations for a practical example for which the underlying distributions are estimated by empirical cdfs. For discussion of cases where one of the distributions is known and the other unknown, see Jamalzadeh [2010].

We compare five Normal $N(\mu, \sigma^2)$ distributions to the standard Normal $N(0, 1)$ distribution, as baseline. Figure 2 displays them. The parameter choices are the baseline itself, and the $N(0.2, 1)$, $N(1, 1)$, $N(0, 2)$, $N(3, 2)$, and $N(5, 1.5)$ distributions. These give a range of parameter choices for varying mean and standard deviation, and for a range of classical ES. The comparison between $N(0, 1)$ and $N(0.2, 1)$ corresponds to Cohen's $d = 0.2$, a *small* ES. We include a check on the baseline comparison with itself.

Table 1 shows the mean and standard deviation of Monte Carlo simulations of Cohen's d, Cliff's δ, the QAD and QCES, and the KL divergence, for Normal distribution comparisons with a $N(0, 1)$ baseline. For each paired comparison we obtain a random sample of 100 observations from each distribution and then compute ES for the comparison. We repeat this 10000 times and compute the mean and standard deviation for each ES for each simulation. All the ES increase as the mean is shifted. For the $N(0, 2)$ comparison, i.e. with the same mean and larger scale than baseline, Cohen's d and Cliff's δ ES remain close to 0 (no

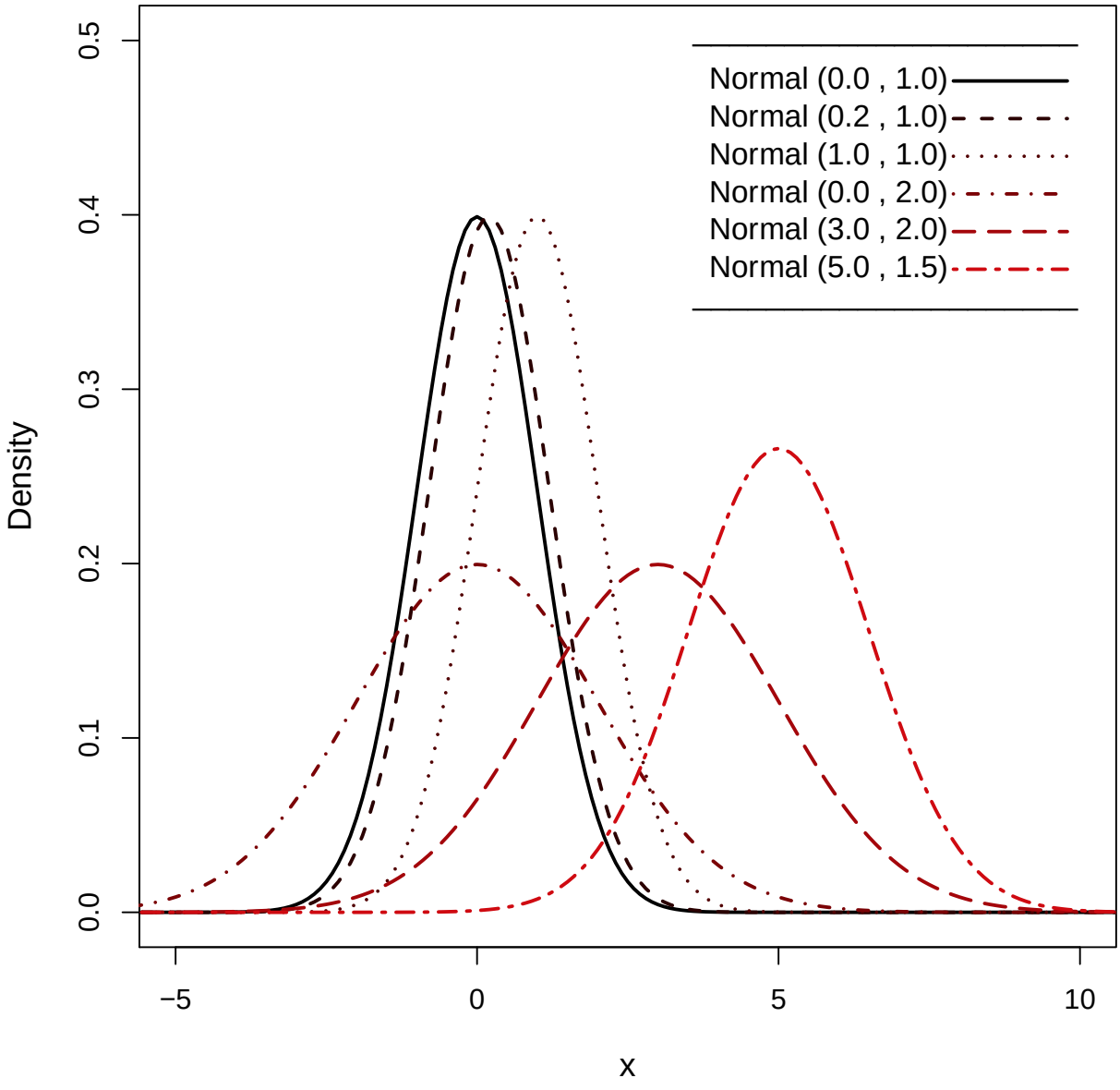

Figure 2: The pdfs for the compared Normal distributions.

Table 1: Mean and standard deviation of Monte Carlo simulations of Cohen's d, Cliff's δ, the QAD and QCES, and the KL divergence, for Normal distribution comparisons with $N(0, 1)$ baseline.

		Effect Size				
		Cohen's d	Cliff's δ	KL	QAD	QCES
$N(\ 0\ , 1.0)$	mean	-0.002	-0.005	0.133	0.135	0.071
	sd	(0.141)	(0.078)	(0.074)	(0.077)	(0.045)
$N(0.2, 1.0)$	mean	0.200	0.110	0.199	0.220	0.122
	sd	(0.144)	(0.076)	(0.101)	(0.119)	(0.068)
$N(1.0, 1.0)$	mean	1.007	0.521	0.929	0.996	0.521
	sd	(0.149)	(0.061)	(0.198)	(0.144)	(0.066)
$N(0.0, 2.0)$	mean	-0.002	-0.002	1.354	0.810	0.214
	sd	(0.142)	(0.085)	(0.363)	(0.127)	(0.028)
$N(3.0, 2.0)$	mean	1.913	0.819	6.022	2.992	0.821
	sd	(0.186)	(0.041)	(1.180)	(0.222)	(0.041)
$N(5.0, 1.5)$	mean	3.940	0.994	13.256	5.002	0.994
	sd	(0.252)	(0.005)	(2.226)	(0.182)	(0.003)

effect), but QCES=0.214 which would suggest a *medium* effect. This illustrates that Cohen's d and d_r are not particularly sensitive to scale changes. The unbounded measures — Cohen's d, KL, and the QAD — have larger standard deviations as the value of the ES increases. The bounded measures — Cliff's δ and the QCES — tend to have smaller standard deviations for larger ES. When one distribution is dominant, this is reflected in both Cliff's δ and the QCES having approximately similar values. As some of these ES are skewed, we also computed the medians and approximate 95% probability intervals on the median for each ES. These gave a similar interpretation.

Figure 3 shows the quantile function (2) for these distributions, with the solid line representing the quantile function of the baseline $N(0, 1)$. Consider distribution functions F and G with means μ_F, μ_G and standard deviations σ_F, σ_G. As the mean increases, the quantile function shifts upwards. In the case of $\sigma_F = \sigma_G$, this shift is the increase in the mean. QAD(F, G) is the area between the quantile functions for F and G. For Normal distributions with equal spread, QAD(F , G) $= |\mu_F - \mu_G|$. In the case where $\mu_F = \mu_G$ and $\sigma_F < \sigma_G$ the quantile functions intersect, with the quantile function for G distorted to lower values at lower probabilities and higher values at higher probabilities, indicating heavier tails. Where $\mu_F \neq \mu_G$ and $\sigma_F \neq \sigma_G$, there is typically both shift and distortion. In contrast to ES such as Cohen's d, the QAD is sensitive to differences across the full distributions and so produces a larger value for the QAD when the standard deviation changes but the mean parameter stays fixed.

Figure 4 plots the vertical quantile comparison functions (3) between distributions F and G, using F as baseline in the left-hand panel, and — reflected — G as baseline in the right-hand panel. $F = G$ corresponds to the solid line $y = x$ in both panels. When two distributions are close, $V_F^G(p)$ and $V_G^F(p)$ are functions close to $y = x$. Specific types of departure from the solid line suggest potential orderings of the two distributions Barlow and Proschan [1975]. Intersections indicate that the distributions may have close mean parameters and different standard deviations. A function lying strictly above or below the solid line indicates that one of the distributions stochastically dominates the other. When one distribution is very far from the other, the vertical quantile comparison function is far from the solid line. The QCES is computed using the area(s) between the curve and the solid line. Thus, the more dissimilar the distributions, the larger the QCES.

Figure 5 illustrates how the value of the QAD (4) changes as we vary the mean and standard deviation. Large differences in mean correspond to larger ES. For fixed means, large differences in the scale parameter lead to larger ES. Differences in both location and scale lead to even larger ES. An equivalent graph for the QCES (7) is illustrated in Figure 6. This ES is always in (0,1), with higher values representing large

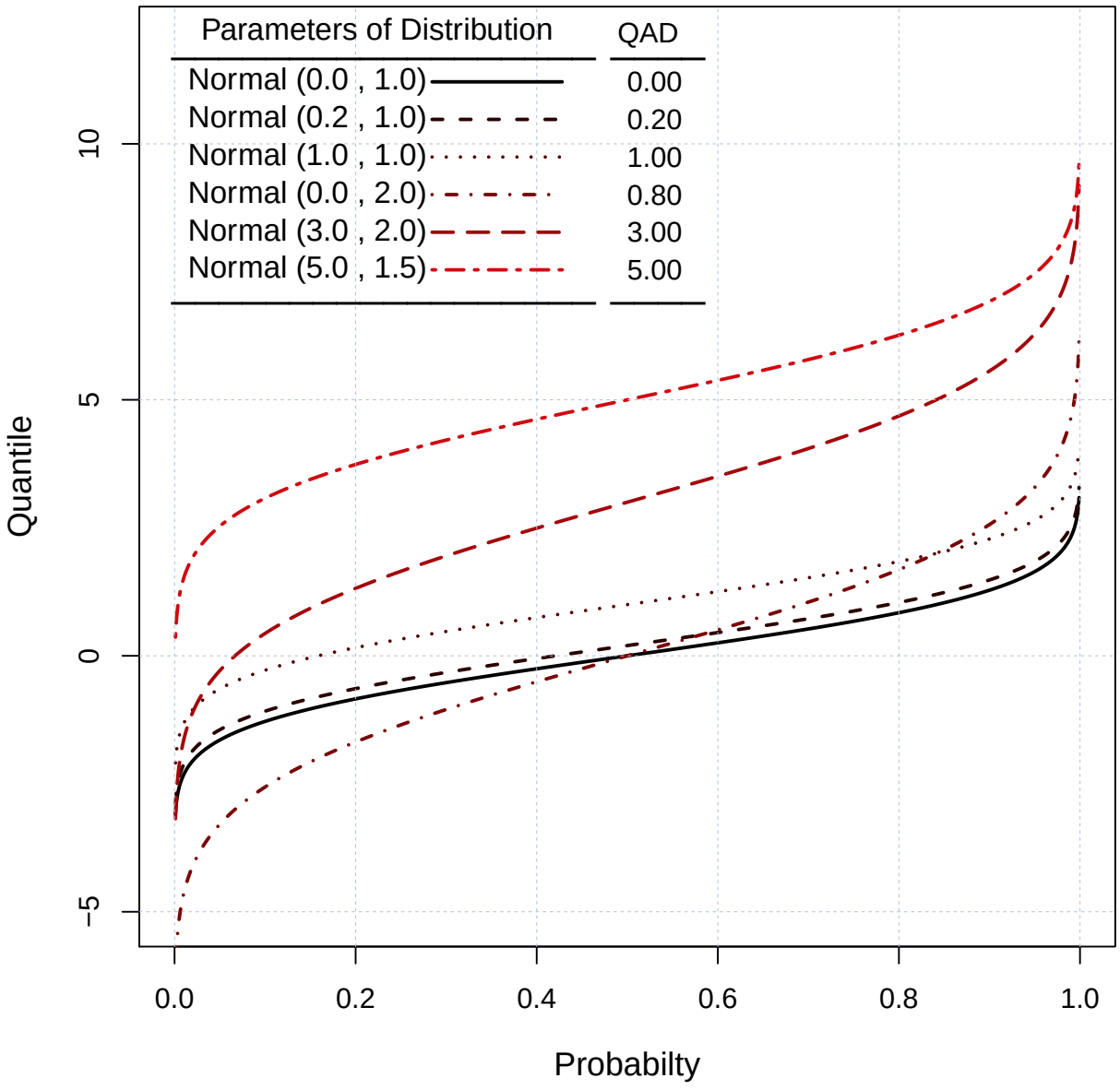

Figure 3: The quantile function (2) for some Normal distributions, with associated QAD, where the Normal distribution $N(0, 1)$ is the baseline.

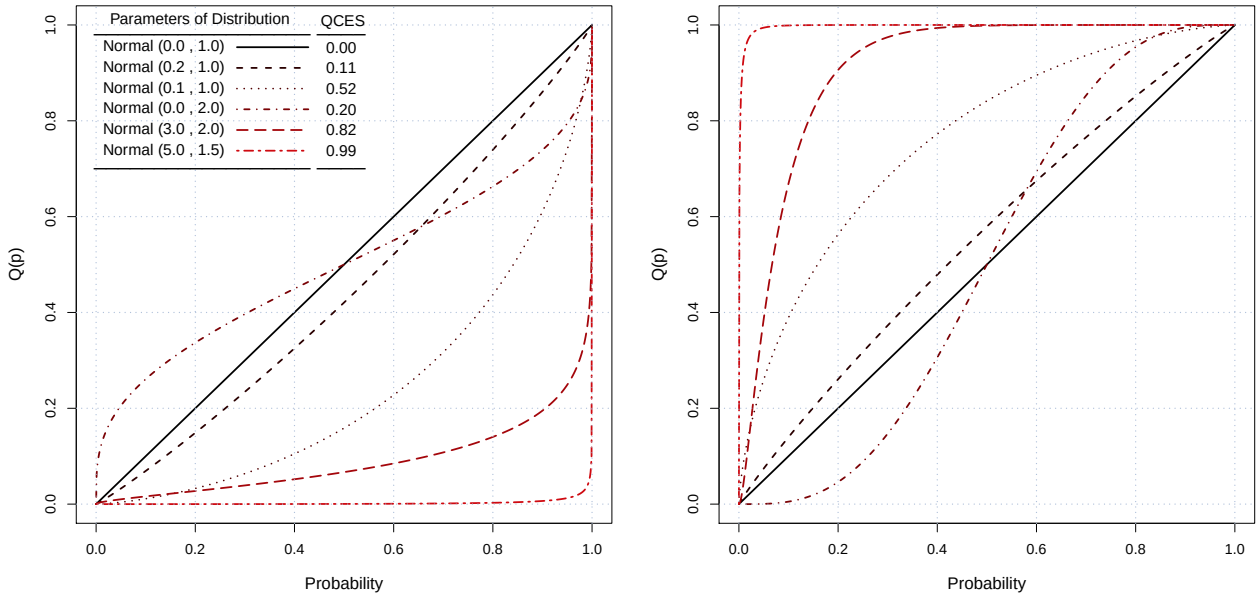

Figure 4: The vertical comparison quantile function (3) for some Normal distributions, with associated QCES. The left panel shows $V_F^G(p) = G(F^{-1}(p))$. The right panel shows the corresponding function $V_G^F(p) = F(G^{-1}(p))$.

effects. Note that because this measure highly depends on the degree of overlap of distribution functions, the ES will not change substantially when one of the distributions has a large standard deviation relative to the other.

5 Two-sample Weibull distribution comparisons: simulation study

In this section we investigate the behaviour of these ES for the Weibull distribution Weibull [1951], which is used to describe the statistical behaviour of many phenomena due to its flexibility. The shifted Weibull distribution Johnson et al. [1994] for a random variable $X \sim W(\alpha, \lambda, \theta)$ has pdf:

$$f(x; \alpha, \lambda, \theta) = \frac{\alpha}{\lambda}\left(\frac{x-\theta}{\lambda}\right)^{\alpha-1} \exp\left\{-\left(\frac{x-\theta}{\lambda}\right)^{\alpha}\right\} \qquad x \geq \theta, \tag{12}$$

where $\alpha > 0$ is a shape parameter, $\lambda > 0$ is a scale parameter, and θ is the threshold or location parameter. When $\theta = 0$, this function reduces to the two-parameter distribution, and we use the notation $X \sim W(\alpha, \lambda)$.

We compare Weibull distributions with different shape and scale parameters to a baseline $W(1, 1)$ distribution, equivalent to an exponential $Exp(1)$ distribution. We choose counterpart Weibull distribution parameters in such a way so as to cover the variety of shapes the distribution may take, as shown in Figure 7.

As for the Normal distribution comparison, we carried out a Monte Carlo exploration of ES for Weibull distribution comparisons with a $W(1, 1)$ baseline. Means and standard deviations for simulated ES are given in Table 2. Small changes in the shape parameter produce small ES for all measures. Larger changes in shape parameter have only a small impact on Cohen's d and Cliff's δ, suggesting that these do not capture such changes well. The QCES for the $W(1.7, 1)$ and $W(0.5, 1)$ comparisons to baseline indicate that these changes in shape and scale have a larger practical effect than Cohen's d and Cliff's δ suggest.

Figure 8 shows the corresponding quantile functions, with the solid line representing the quantile function for the baseline $W(1, 1)$ distribution. As the scale parameter λ increases, the quantile function shifts upwards. Different shape parameters lead to distorted quantile functions which intersect the quantile function for the baseline $W(1, 1)$ distribution. The magnitude of the QAD effect size (4) is the area between two such quantile functions. In general, larger scale parameters produce larger ES. Figure 10

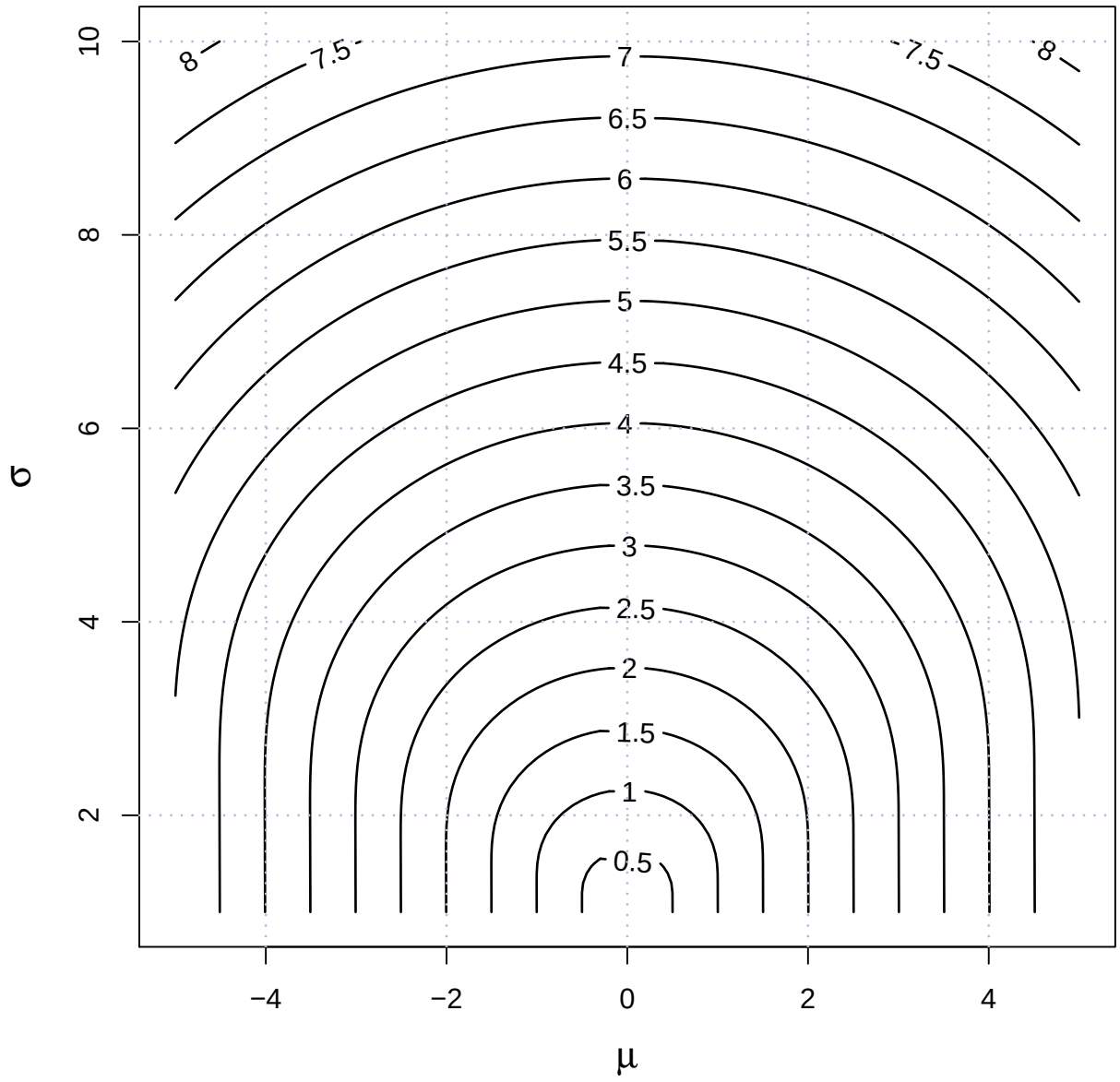

Figure 5: Contour plot of the QAD (4) for Normal distribution comparisons with $N(0,1)$ baseline: changes in ES as we vary μ and σ. The point (0,1) locates the baseline. ES are positive unbounded.

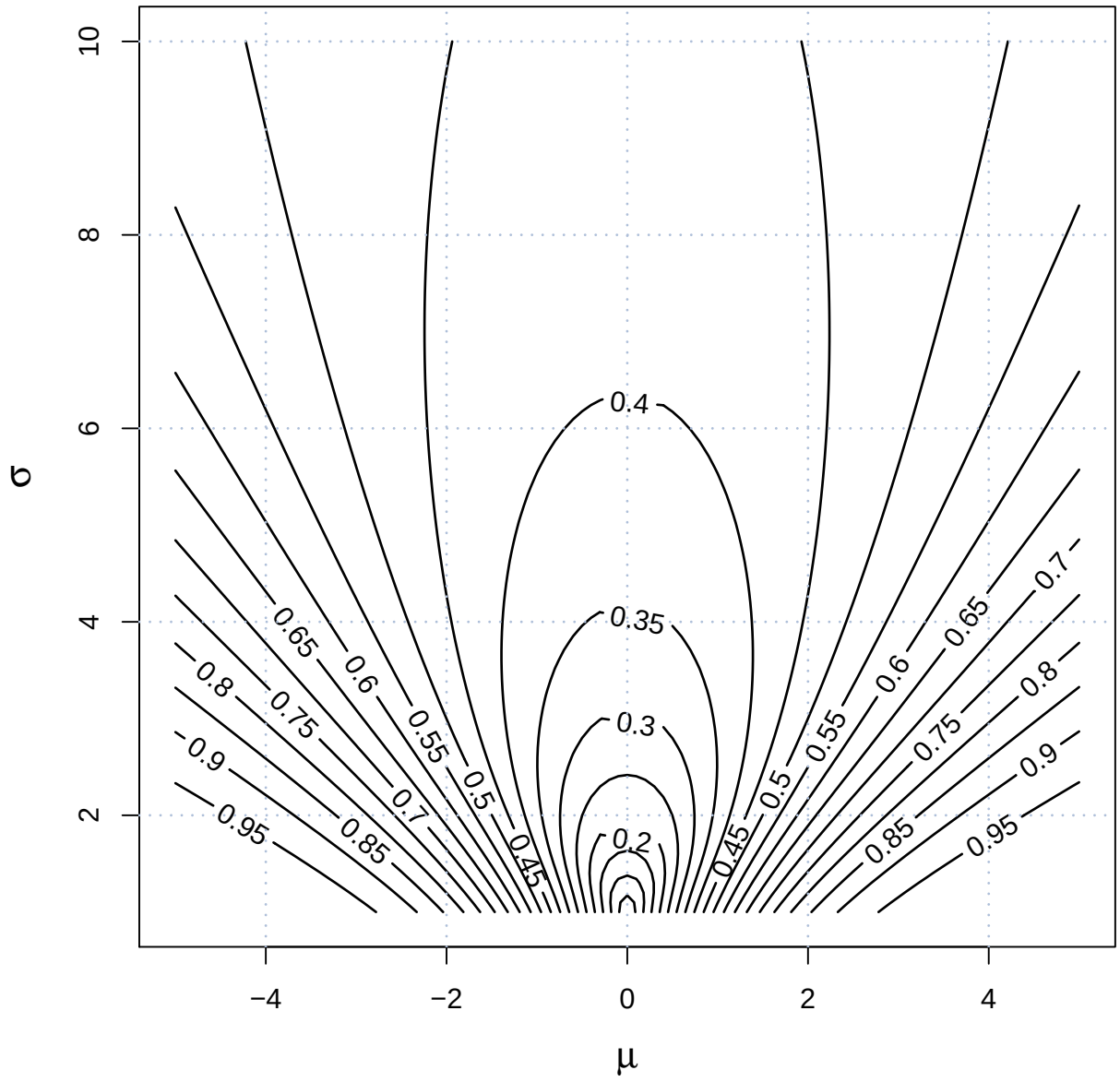

Figure 6: Contour plot of the QCES (7) for Normal distribution comparisons with $N(0,1)$ baseline: changes in ES as we vary μ and σ. The point (0,1) locates the baseline. ES are in (0,1) by design.

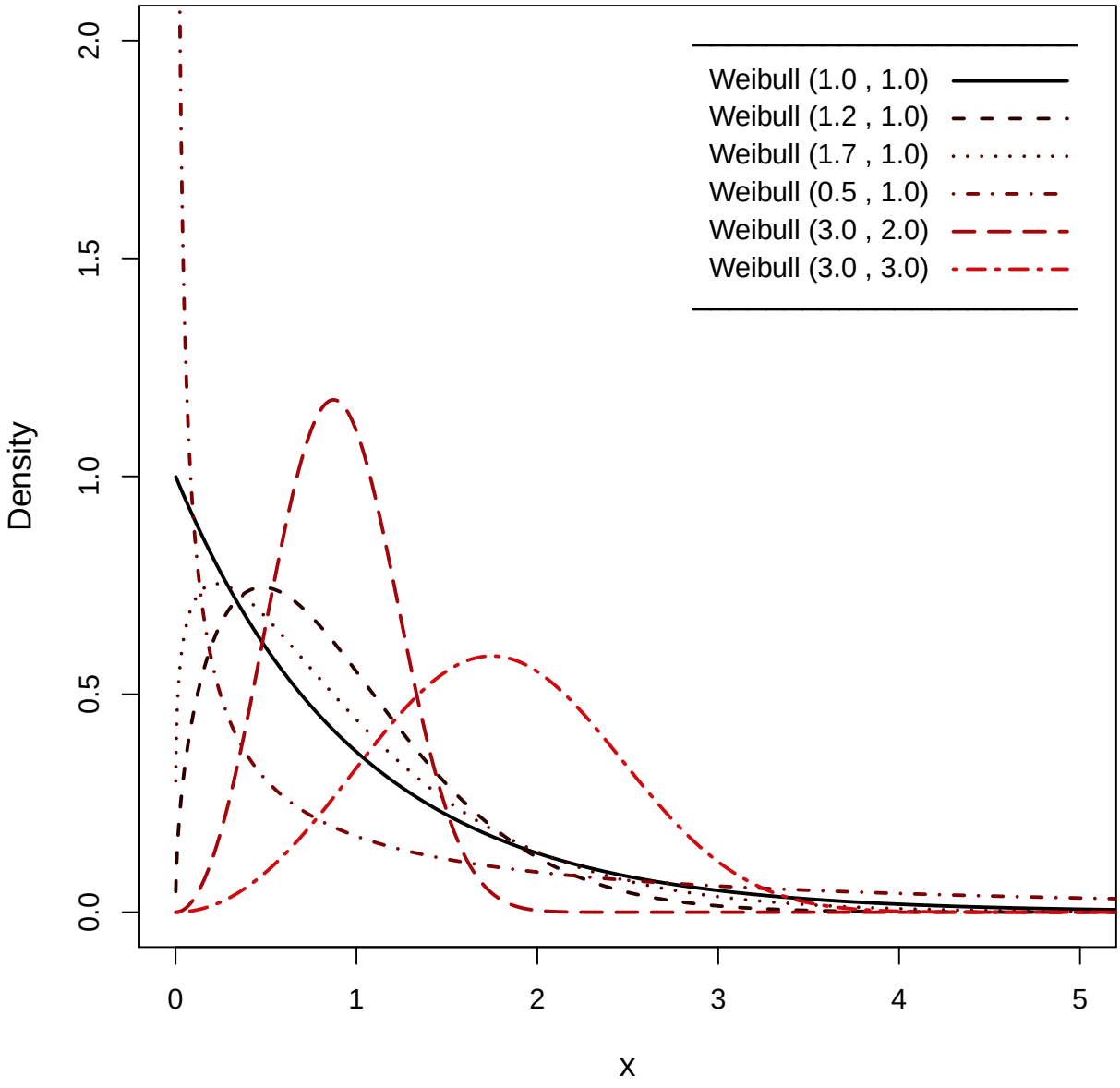

Figure 7: The pdfs for the compared Weibull distributions.

Table 2: Mean and standard deviation of Monte Carlo simulations of Cohen's d, Cliff's δ, the QAD and QCES, and the KL divergence, for Weibull distribution comparisons with $W(1,1)$ baseline.

		Effect Size				
		Cohen's d	Cliff's δ	KL	QAD	QCES
$W(1.0,1)$	mean	-0.000	0.009	0.120	0.130	0.071
	sd	(0.143)	(0.082)	(0.067)	(0.079)	(0.045)
$W(1.2,1)$	mean	-0.064	0.025	0.154	0.163	0.090
	sd	(0.141)	(0.084)	(0.065)	(0.075)	(0.043)
$W(1.7,1)$	mean	-0.126	0.075	0.349	0.296	0.174
	sd	(0.137)	(0.079)	(0.069)	(0.070)	(0.040)
$W(0.5,1)$	mean	0.322	-0.087	1.185	1.231	0.217
	sd	(0.094)	(0.081)	(0.491)	(0.398)	(0.038)
$W(2.0,3)$	mean	0.953	0.239	0.880	0.878	0.600
	sd	(0.213)	(0.018)	(0.108)	(0.088)	(0.061)
$W(3.0,3)$	mean	1.718	0.289	1.555	1.690	0.788
	sd	(0.232)	(0.002)	(0.242)	(0.130)	(0.046)

shows how the QAD changes as we vary α, λ from the baseline at (1,1). For a fixed value of α, an increase in λ results in a sharp increase in ES, and especially for $\alpha < 1$.

Figure 9 displays the vertical shift quantile functions and the associated QCES (7) for these comparisons. Changing the scale parameter affects the ES considerably; changes in shape have less impact. The contour plot shown in Figure 11 summarises changes in ES as we vary (α, λ) from the baseline point $(1,1)$.

6 Application: analysis of clickstream data

A clickstream is the record of the visiting behaviour and history of a customer to an internet website such as an online retailer. Server log files provide tracking and purchase information for visitors to the website. The data recorded includes the time and date of visit, the IP address, details as to the history of entry to the site, whether the visit resulted in a purchase, and if so the amount of revenue. Analysis of such data enhances understanding and prediction of website visitor behaviour Andersen et al. [2000], usually with the aim of maximising customer sales and revenue. Session data forms a subset of the record and explains a customer's browsing behaviour during a single visit and is the most commonly analysed aspect of e-commerce data. Key features are: which pages are most frequently viewed, the average view time per page, average length of a path through a site, common entry and exit points, and other aggregate measures. Exploratory analysis of such data, even though often superficial, is useful for improving system performance and providing marketing decision-support Markov and Larose [2007]. A simple common question, for example, is whether the revenues extracted from customers arriving from different entry points is about the same or different. This is a simple two-sample comparison. The amount of data available is huge, potentially millions of records per day. Therefore, standard methods for exploring such simple hypotheses fail as they inevitably yield tiny p-values. Furthermore, the underlying populations are often highly non-Normal. We may instead employ ES to explore such hypotheses, for which we need the ES methods developed in this paper. We use clickstream data from a commercial website. For this illustration, we use data collected for one website for one week in the summer of 2008. Some data cleaning is required in order to remove duplicates and so forth. In all, 10091 visit details were recorded.

For this illustration, we examine session time duration, T, the time a customer spends on the website. We exclude customers who visit only one page. Figure 12 shows histograms of T, right censoring at 100 minutes, with visits classified as leading to sales, T_1, or no sale, T_0. The distributions are visually quite different. Summary statistics suggest larger values of T for visits which lead to sales. In testing whether the means of

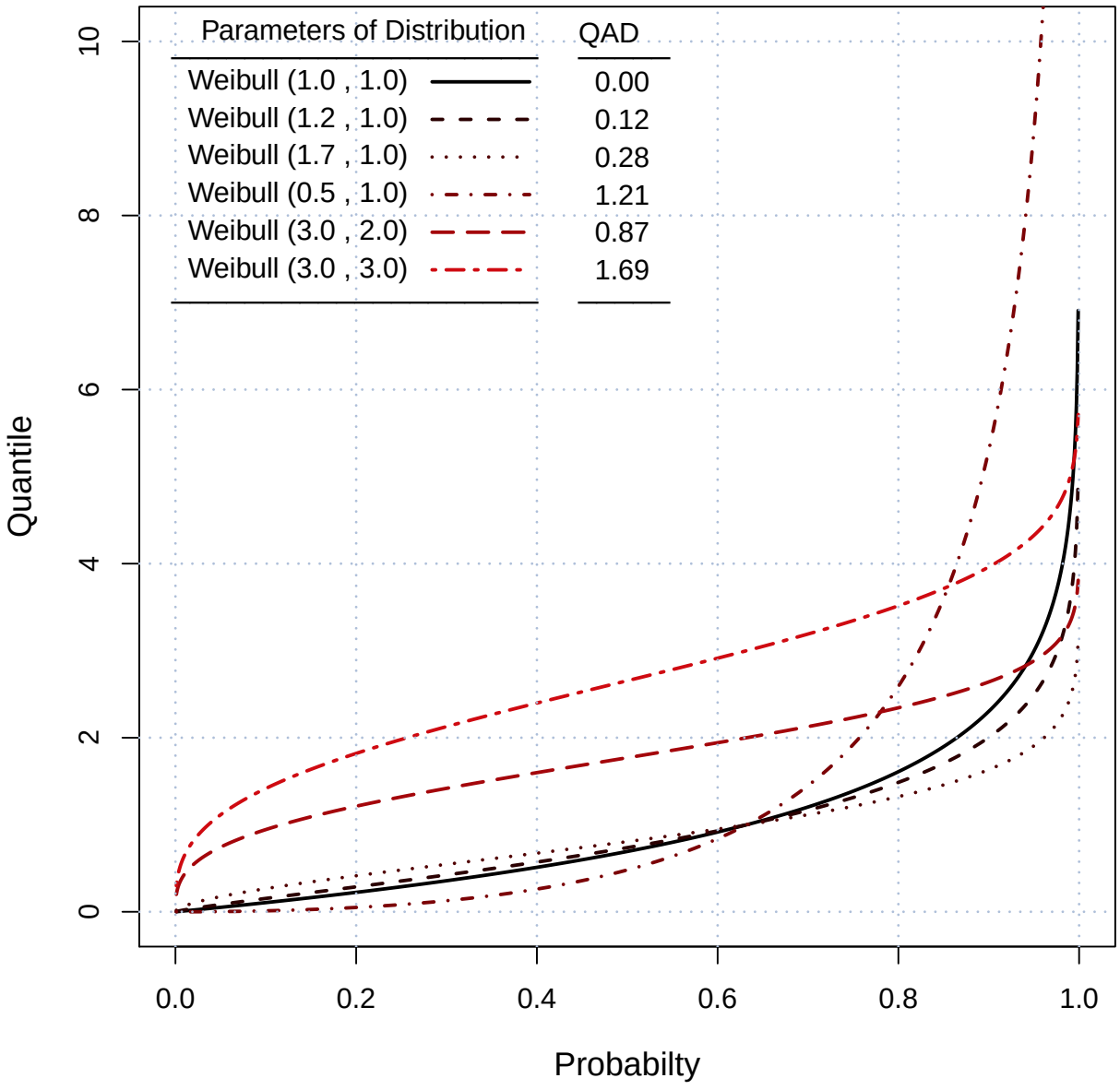

Figure 8: The quantile function (2) for some Weibull distributions, with associated QAD comparing to a baseline $W(1,1)$ distribution.

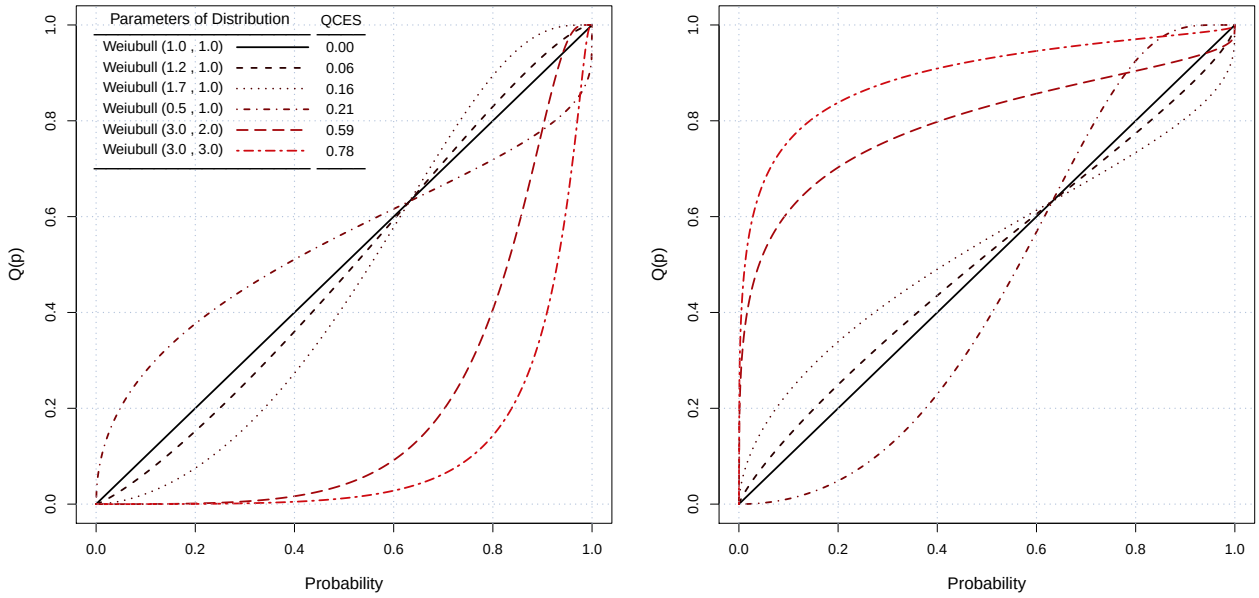

Figure 9: The vertical comparison quantile function (3) for some Weibull distributions, with associated QCES comparing to a baseline $W(1,1)$ distribution. The left panel shows $V_F^G(p) = G(F^{-1}(p))$. The right panel shows the corresponding function $V_G^F(p) = F(G^{-1}(p))$.

the two populations T_0 and T_1 are equal, the standard t-test is not much help. The distributions seem clearly non-Normal, and the sample size is so large that tiny p-values necessarily result. The Mann-Whitney test has similar drawbacks. Cliff's nonparametric ES is $\delta = 0.22$, implying some degree of dominance. Cohen's $d = 0.95$ suggests a large effect under an assumption of Normality, but this is inappropriate here. We thus compute the quantile-based QAD and QCES effect sizes. We may compute these either by using empirical cdfs or by fitting suitable distributions to each set of observations.

6.1 Fitting a distribution and then computing effect sizes

We fit separate Weibull distributions to T_0 and T_1 via maximum likelihood. The fits are shown in Figure 13. The shaded area is the overlap in the two distributions. The shape parameters are $\alpha_1 = 1.39 > 1$ and $\alpha_0 = 0.76 < 1$ for the sales and non-sales groups respectively. The sales-group distribution also has a larger scale parameter. Quantile plots suggested that these fits were good except in the furthest extremities. The quantile functions for these distributions are shown in Figure 14. The shaded area is the QAD (4), which evaluates to $\mathrm{QAD}(T_1\,,\,T_0) = 9.35$, suggesting that on average the quantiles of T for the sales group exceeds quantiles in the non-sales group by 9.35 minutes. It should be noted that the QAD and QCES do tell us which distribution is dominant, when a dominance exists.

Figure 15 depicts the vertical quantile comparison functions (3) for these distributions. The QCES is the area between these functions and turns out in this case to be QCES$= 0.61$. Following the argument in §3.3, we would judge this as indicating a very strong ES.

In order to assess sensitivity, we take the $W(0.76, 4.95)$ distribution for T_0 as baseline and calculate the QCES when we vary the fitted parameters of the distribution for T_1. Figure 16 shows the resultant contour plot. The baseline is indicated by a black square. The actual QCES is indicated by a shaded circle. The plot suggests that the sales group has fitted parameters which are very far from parameter values which would produce a much smaller ES.

6.2 Using empirical quantile functions to generate effect sizes

We may instead use the empirical quantile functions, derived from the empirical cdfs, to compute ES. This avoids the need to estimate parametric forms for the variables under study. Using the empirical

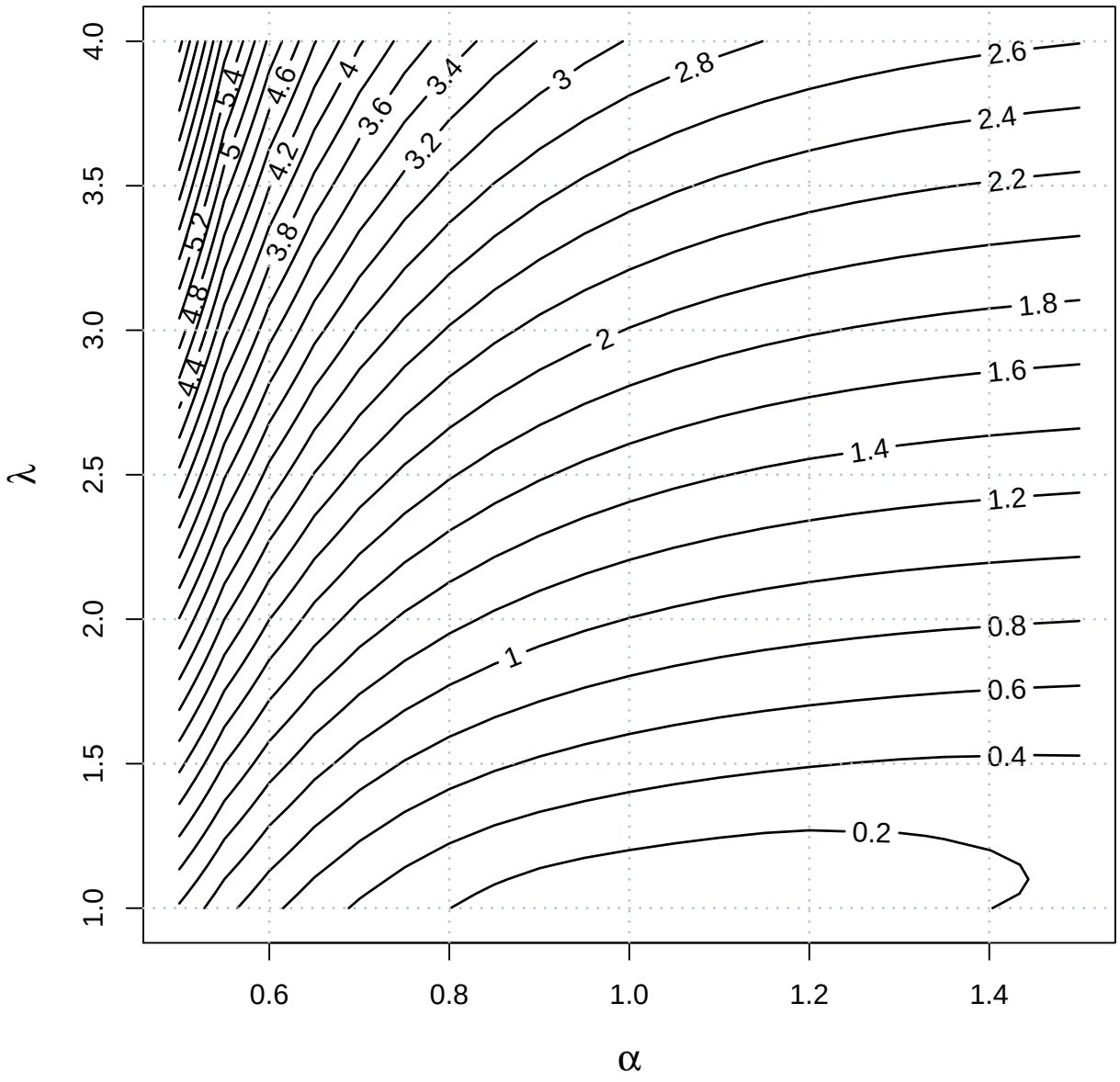

Figure 10: Contour plot of the QAD (4) for Weibull distribution comparisons with $W(1,1)$ baseline: changes in ES as we vary α, λ. The point (1,1) locates the baseline. ES are positive unbounded.

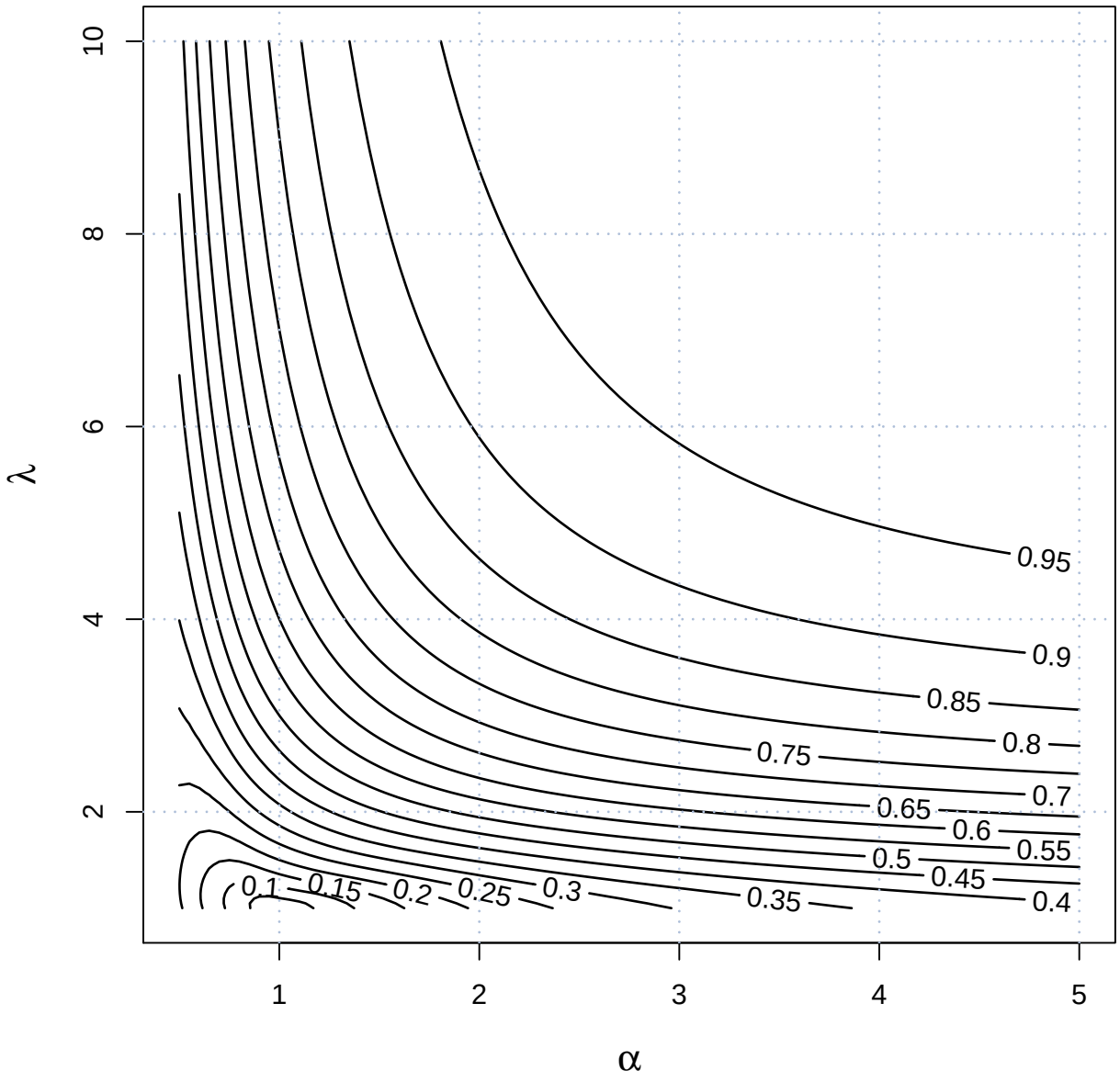

Figure 11: Contour plot of the QCES (7) for Weibull distribution comparisons with $W(1,1)$ baseline: changes in ES as we vary α, λ. The point (1,1) locates the baseline. ES are in (0,1) by design.

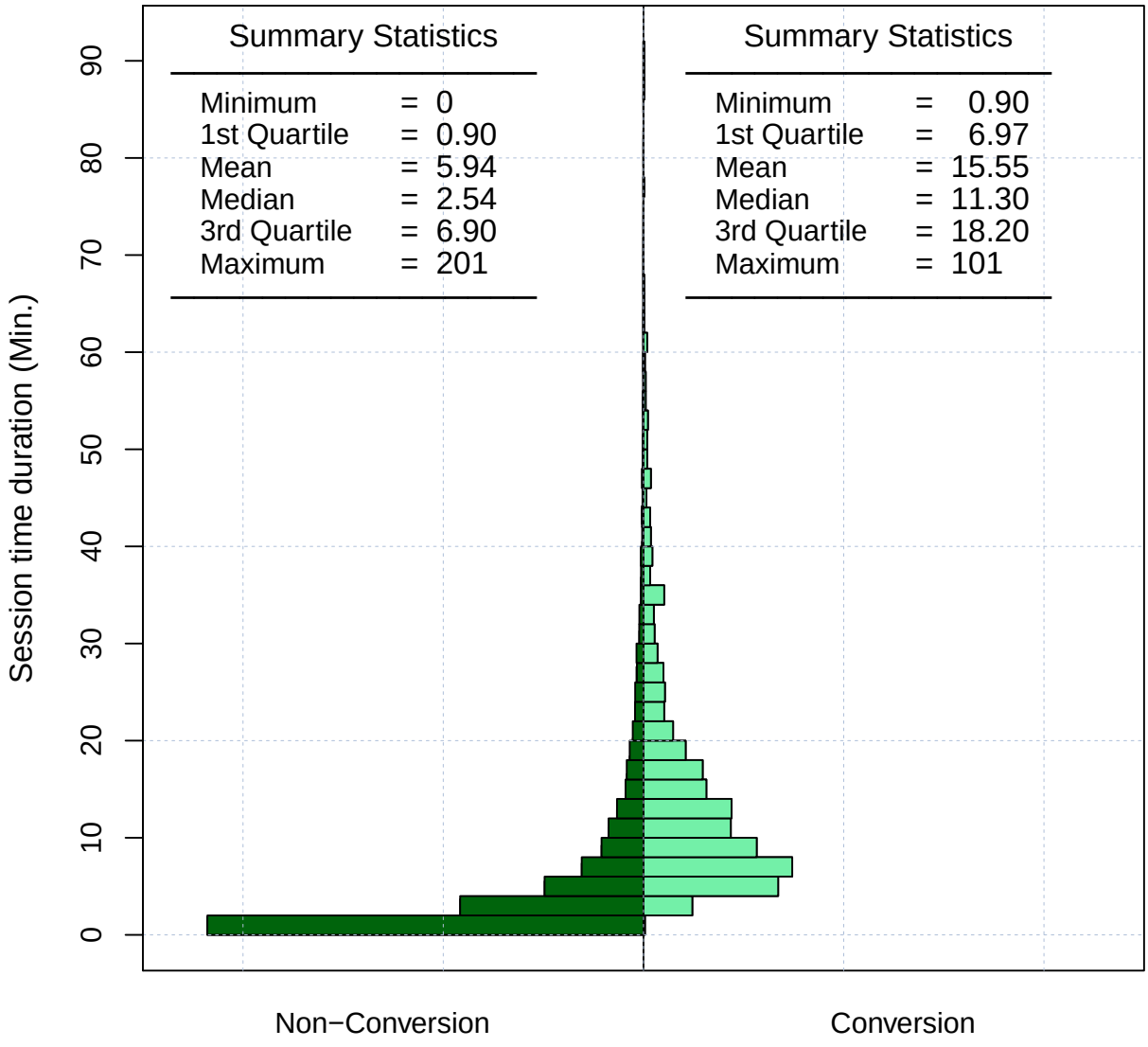

Figure 12: Back-to-back histograms of session time duration for 1,353 website visits resulting in a sale, and durations for 8,747 non-sale visits.

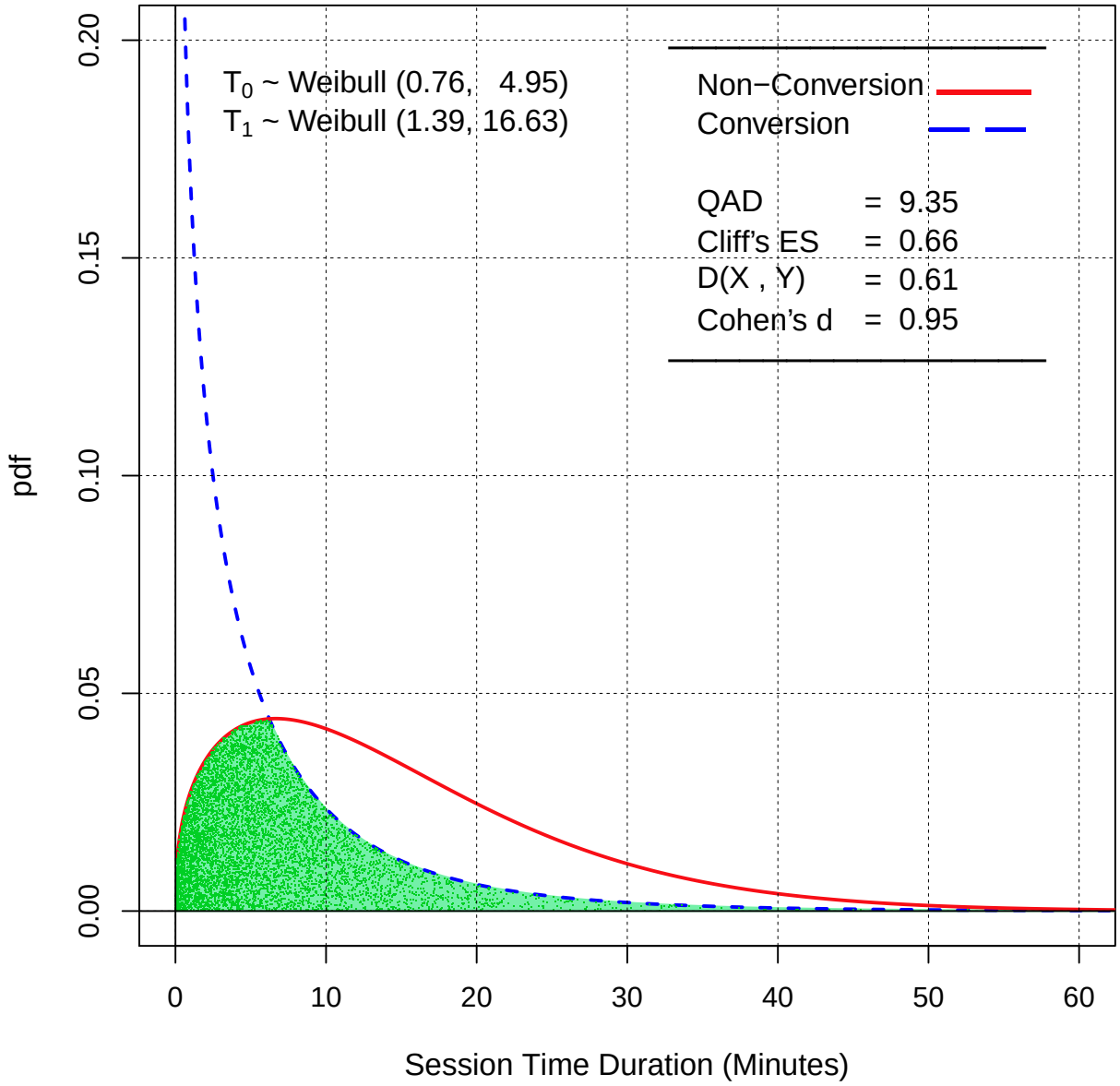

Figure 13: Fitted pdfs of session time duration T, separately for sales and non-sales visits, estimation via maximum likelihood.

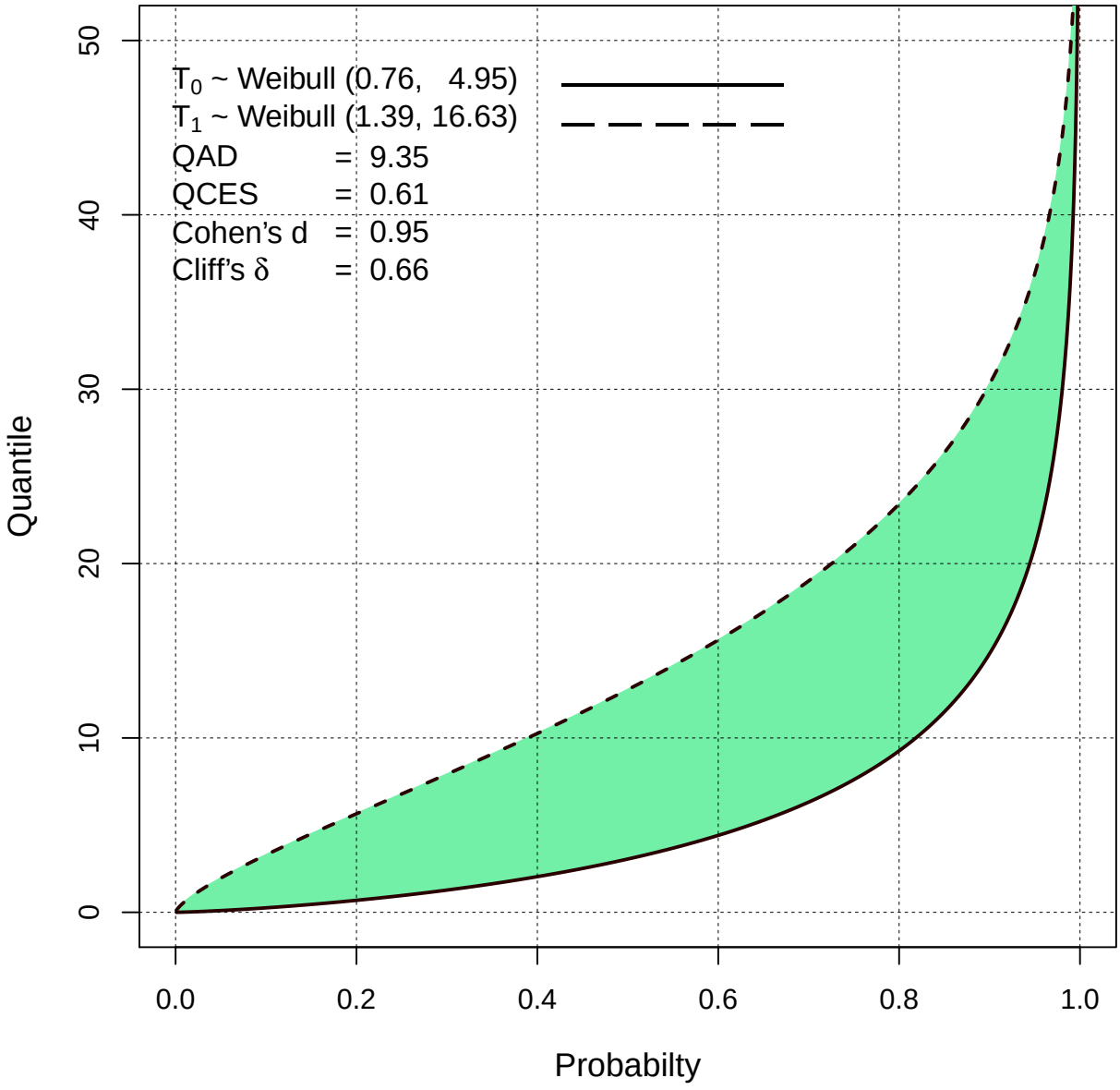

Figure 14: The fitted quantile functions for session time duration, separately for sales and non-sales visits.

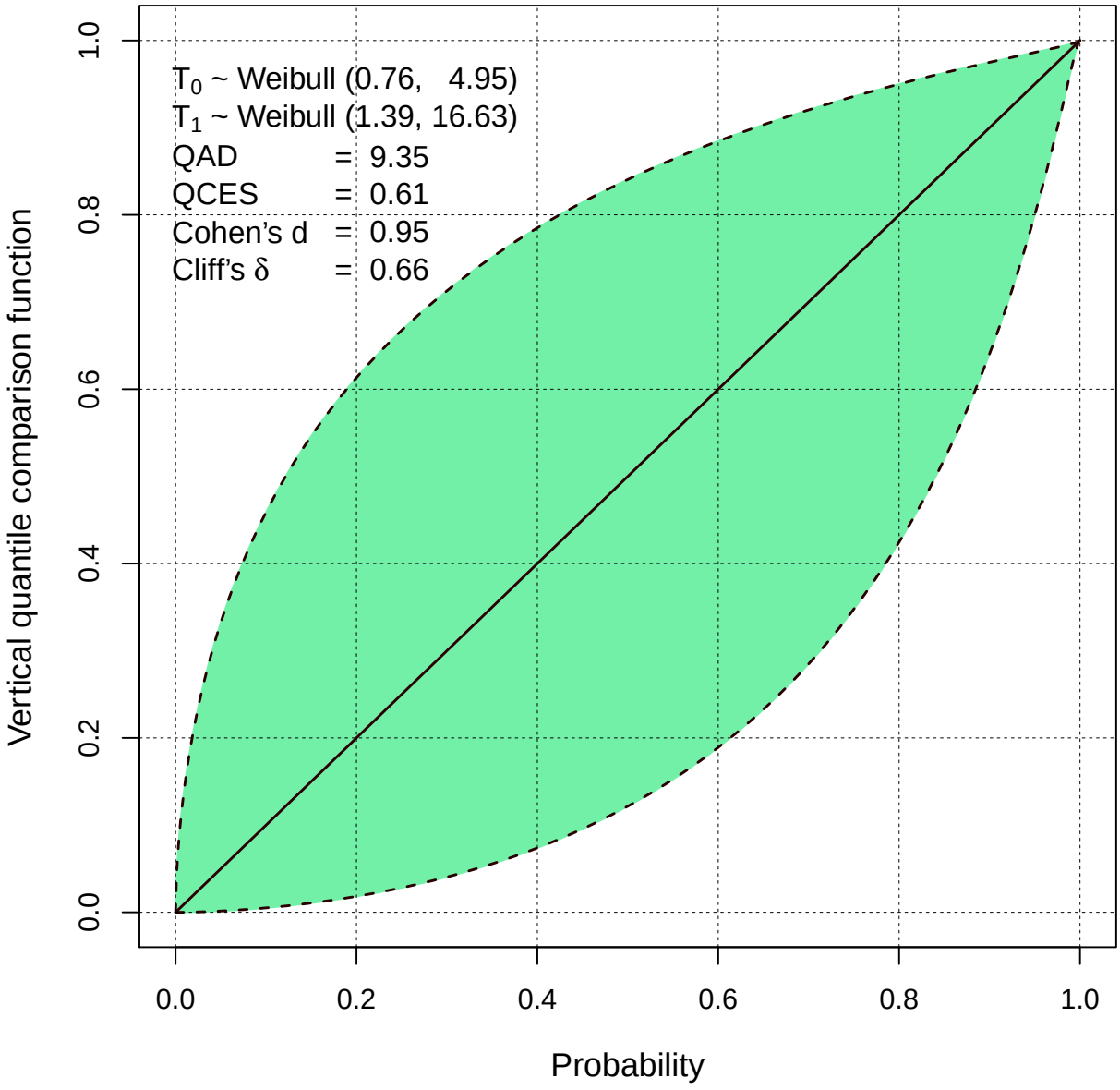

Figure 15: Vertical comparison quantile functions for T_0 and T_1.

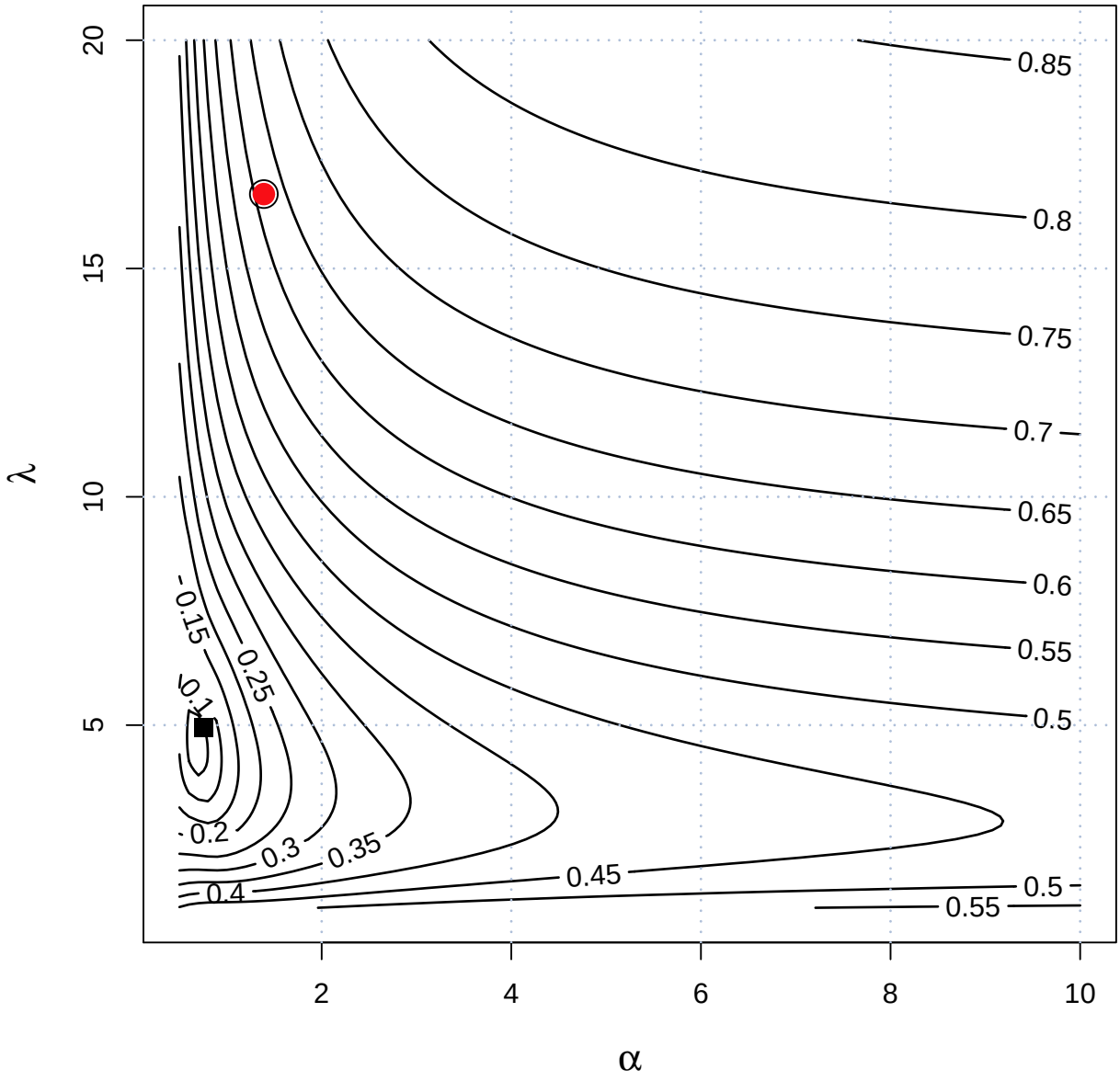

Figure 16: Sensitivity plot: a contour plot of QCES values for varied parameter choices for sales-group distribution, comparing to baseline distribution T_0.

cdfs, we computed QAD=9.08 which is slightly smaller than the corresponding parametric estimate. The corresponding value of the QCES is 0.654, again suggesting a very large effect size.

7 Inference based on bootstrapping

The theoretical distributions of the QAD and the QCES are complicated to find. Thus, we derive a bootstrap approximation of their distributions given the observed samples Davison and Hinkley [2006]. For the two-sample comparison, repeated resamples are drawn with replacement from the two sets of observations and ES and so forth calculated from the resamples. We repeated this resampling process 10,000 times in order to find bootstrap distributions and confidence intervals for our ES Hesterberg et al. [2003]. We did this for the parametric method described in §6.1, which involves fitting Weibull distributions to each resample and then computing ES, and for the nonparametric method described in §6.2, which generates a different empirical cdf for each resample. Table 3 shows the result of the bootstrap sampling for QAD and QCES for both parametric and non-parametric approaches. F_0 and F_1 are the fitted cdfs for T_0 and T_1 respectively, and $\widetilde{F}_0$ and $\widetilde{F}_1$ are the corresponding empirical cdfs. The column entitled *Observed* gives the ES computed from the original sample. The remaining columns summarise the bootstrapped ES. Histograms of the bootstrapped ES are all reasonably Normal in shape, and the summaries give no cause for concern.

Table 3: Bootstrap summary statistics based on 10,000 resamples: estimation of the standard error, 95% confidence interval, and bias for each ES, calculatd for parametric (P) and empirical (E) cases.

	Effect Size	Observed	Mean	SE	95% CI	Bias
P	QAD QAD(F_0 , F_1)	9.346	9.328	0.346	(8.65 , 9.99)	+0.018
E	QAD QAD($\widetilde{F}_0$, $\widetilde{F}_1$)	9.081	9.056	0.347	(8.39 , 9.74)	+0.025
P	QCES QCES(F_0 , F_1)	0.611	0.610	0.009	(0.59 , 0.63)	+0.001
E	QCES QCES($\widetilde{F}_0$, $\widetilde{F}_1$)	0.645	0.646	0.009	(0.63 , 0.66)	−0.001

8 Conclusion

Commonly used ES for two-sample comparison studies are mostly based on Normality assumptions and limited aspects of changes in parameter. In this paper we have introduced two ES measures, the QAD and the QCES, which are based on quantile functions and which better summarise differences between distributions over the full range of probabilities. The QAD is an ES for which differences between distributions are summarised in terms of the original units of measurement. The QCES has been developed as a bounded standardized divergence measure for circumstances where the unit of measurement is not meaningful or relevant. We have investigated these ES for two parametric families and in a practical application, and suggested some heuristic thresholds for what constitutes small and large effects. See Jamalzadeh [2010] for a Bayesian approach for our practical application to clickstream data, and investigation of computation of the QAD and the QCES within the context of prior and posterior distributions.

References

Algina, J., H. Keselman, and R. Penfield (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods 10*, 317–328.

Andersen, J., R. Larsen, A. Giversen, T. Pedersen, A. Jensen, and J. Skyt (2000). Analyzing clickstreams using subsessions. Technical report tr-00-5001, Department of Computer Science, Aalborg University.

Anderson, C. L. and J. C. Berry (2009). On a simple measure of dominance. *Journal of Statistical Planning and Inference 139*, 1098–1108.

Barlow, R. E. and F. Proschan (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. New York: Holt, Reinhart and Winston.

Cahan, S. and E. Gamliel (2011). First among others? Cohen's d vs. alternative standardized mean group difference measures. *Practical Assessment, Research & Evaluation 16*(10), 1–6.

Cleveland, W. S. (1994). *The elements of graphing data*. Murray Hill, N.J: AT&T Bell Laboratories.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*, 494–509.

Coe, R. (2002, September). It?s the effect size, stupid. what effect size is and why it is important. In *Proceedings of the the Annual Conference of the British Educational Research Association*. University of Exeter, England.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (Revised Edition ed.). Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin 112*, 112–159.

Darlington, R. B. (1973). Comparing two groups by simple graphs. *Psychological Bulletin 79*(2), 110–116.

Davison, A. and D. Hinkley (2006). *Bootstrap Methods and their Application* (8th Edition ed.). Cambridge.

Descôteaux, J. (2007). Statistical power: An historical introduction. *Tutorials in Quantitative Methods for Psychology 3*(2), 28–34.

Doksum, K. A. (1977). Some graphical methods in statistics. a review and some extensions. *Statistica Neerlandica 31*, 53–68.

Doksum, K. A. and A. Samarov (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Annals of Statistics 23*, 1443–1473.

Doksum, K. A. and G. L. Sievers (1976). plotting with confidence - graphical comparisons of 2 populations. *Biometrika 63*, 421–434.

Fleming, T. R., J. R. O'Fallon, P. C. O' Brien, and D. P. Harrington (1980). Modified kolmogorov-smirnov test procedures with applications to arbitrarily right-censored data. *Biometrics 36*, 607–625.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren and C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, pp. 311–339. Hillsdale, NJ: LEA.

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. Chapman and Hall.

Glass, G., B. McGaw, and M. Smith (1981). *Meta-analysis in social research*. Beverly Hills, CA: SAGE Publications.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology 79*(2), 314–316.

Grissom, R. J. and J. J. Kim (2005). *Effect Size for Researches: A Broad Practical Approach*. Psychology Press, Taylor and Francis Group.

Hedges, L. V. and L. Friedman (1993). Gender differences in variability in intellectual abilities: A reanalysis of feingold's results. *Review of Educational Research 63*(1), 94–105.

Hess, M. R. and J. D. Kromrey (2004, April). Robust confidence intervals for effect sizes: A comparative study of Cohens d and cliffs delta under non?normality and heterogeneous variances. In *Proceedings of annual meeting of the American Educational Research Association*. San Diego.

Hesterberg, T., S. Monaghan, D. Moore, A. Clipson, , and R. Epstein (2003). *The Practice of Business Statistics*. W. H. Freeman and Company.

Hinkley, D. (1969). On the ratio of two correlated normal random variables. *Biometrika 56*(3), 635–639.

Hodges, L. and I. Olkin (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

Holmgren, E. C. (1995). The p-p plot as a method of comparing treatment effects. *Journal of the American Statistical Association 90*(429), 360–?365.

Jamalzadeh, A. (2010). *Statistical Analysis of Web Usage Data*. Ph. D. thesis, Mathematical Science Department, Durham University.

Johnson, N., S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions* (2nd ed.), Volume 1. New York: John Wiley and Sons.

Keselman, H., C. Huberty, L. Lix, S. Olejnik, R. Cribbie, B. Donahue, R. Kowalchuk, L. Lowman, M. Petoskey, J. Keselman, and J. Levin (1998). Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses. *Review of Educational Research 68*(3), 350–386.

Krueger, J. (2001). Null hypothesis significance testing. on the survival of a flawed method. *American Psychologist 56*(1), 16–26.

Kulinskaya, E. and R. G. Staudte (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic anova. *British Journal of Mathematical and Statistical Psychology 59*, 97–111.

Kullback, S. (1968). *Information Theory and Statistics*. Publications Inc., Mineola, New York.

Ledesma, R., G. Macbeth, and N. Cortada de Kohan (2009). Computing effect size measures with vista. *Tutorials in Quantitative Methods for Psychology 5*(1), 25–34.

Lenth, R. (2001). Some practical guidelines for effective sample-size determination.

Li, G., R. Tiwari, and M. Wells (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *Journal of the American Statistical Association 91*(434), 689–?698.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Statistics 18*, 50–60.

Markov, Z. and D. Larose (2007). *Data Mining the Web: Uncovering Pattern in Web Content, Structure, and Usage*. Jonh Wiley and Sons.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools 5*(2), 33–?38.

Tukey, J. W. (1960). *A survey of sampling from contaminated normal distributions In I. Olkin et al. (Eds.) Contributions to Probability and Statistics.* HStanford, CA: Stanford University Press.

Weibull, W. (1951, September). A statistical distribution function of wide applicability. *ASME Journal of Applied Mechanic, Transactions of the American Society of Mechanical Engineers 18*(3), 293–297.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Academic Press.

Wilcox, R. R. and T. S. Tian (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *in press 38*(7), 1359–1368.

Wilk, M. B. and R. Gnanadesikan (1968). Probability plotting methods for the analysis of data. *Biometrika 55*, 1–17.

Wilson Van Voorhis, C. and B. Morgan (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology 3*(2), 43–50.

Appendix: Algorithms

Algorithm 1 The empirical quantile function, EQF($x[\,]$,$p[\,]$)

Require: a vector of observations $x[\,]$.
Require: a vector of probabilities $p[\,]$.
$n_x \leftarrow$ length of the vector $x[\,]$
$n_p \leftarrow$ length of the vector $p[\,]$
for $i = 1$ to n_x **do**
  $F_x[i] \leftarrow i/n_x$
end for
for $i = 1$ to n_p **do**
  $q_x[i] \leftarrow$ minimum value of $x[\,]$
end for
for $j = 1$ to n_p **do**
  for $i = 1$ to $n_x - 1$ **do**
    if $p[j] \geq F_x[i]$ **then**
      $q_x[j] \leftarrow x[i + 1]$
    end if
  end for
end for
return $q_x[\,]$

Algorithm 2 The empirical cdf, ECDF($x[\,]$, $q[\,]$)

Require: a vector of observations $x[\,]$.
Require: a vector of quantiles $q[\,]$.
$x[\,] \leftarrow$ sort the vector of $x[\,]$ by ascending order
$n_x \leftarrow$ length of the vector $x[\,]$
$q[\,] \leftarrow$ sort the vector of $q[\,]$ and keep unique values
$n_q \leftarrow$ length of the vector $q[\,]$
for $i = 1$ to n_q **do**
  $sum \leftarrow 0$
  for $j = 1$ to n_q **do**
    if $x[j] \leq q[i]$ **then**
      $sum \leftarrow sum + 1$
    end if
  end for
  $F[i] \leftarrow sum/n_x$
end for
return $F[\,]$

Algorithm 3 The empirical QAD effect size, $\textsc{qad}(x[\,], y[\,])$

Require: a vector of observations $x[\,]$.
Require: a vector of observations $y[\,]$.
$x[\,] \leftarrow$ sort vector of $x[\,]$ by ascending order
$y[\,] \leftarrow$ sort vector of $y[\,]$ by ascending order
$N_x \leftarrow$ length of the vector $x[\,]$
$N_y \leftarrow$ length of the vector $y[\,]$
for $i = 1$ to n_x **do**
    $F_x[i] \leftarrow i/n_x$
end for
for $i = 1$ to n_y **do**
    $F_y[i] \leftarrow i/n_y$
end for
$p[\,] \leftarrow$ merge F_x and F_y and sort them by ascending order
$n_p \leftarrow$ length of the vector $p[\,]$
$Q_x \leftarrow \textsc{eqf}(x[\,], p[\,])$
$Q_y \leftarrow \textsc{eqf}(y[\,], p[\,])$
for $i = 1$ to n_p-1 **do**
    $d_p[i] \leftarrow p[i+1] - p[i]$
end for
for $i = 2$ to n_p **do**
    $d_Q[i-1] \leftarrow \left|Q_x[i] - Q_y[i]\right|$
end for
$qad \leftarrow 0$
for $i = 2$ to n_p-1 **do**
    $qad \leftarrow qad + d_Q[i] \times d_p[i])$
end for
return qad

Algorithm 4 The area of a non-self-intersecting polygon with n vertices, $\textsc{area}(v[\,,\,])$.

Require: a $(n+1) \times 2$ matrix of vertices $v[\,,\,]$ of a non-self-intersecting polygon.
$area \leftarrow 0$
$n_v \leftarrow$ the number of vertices
for $i = 1$ to n_v **do**
    $area \leftarrow area + \frac{1}{2} \times \big(v[i,1] \times v[i+1,2] - v[i,2] \times v[i+1,1]\big)$
end for
return $area$

Algorithm 5 The empirical divergence measure, $\textsc{div}(x[\,], y[\,])$

Require: a vector of observations $x[\,]$.
Require: a vector of observations $y[\,]$.
$x[\,] \leftarrow$ sort the vector of $x[\,]$ by ascending order
$GGinv \leftarrow [0, \textsc{ecdf}(x[\,], x[\,]), 1]$
$FGinv \leftarrow [0, \textsc{ecdf}(y[\,], x[\,]), 1]$
$n \leftarrow$ length of GGinv
for $i = 1$ to n **do**
  $v[i, 1] \leftarrow GGinv[i]$
end for
for $i = 1$ to n **do**
  $v[i, 2] \leftarrow FGinv[i]$
end for
for $j = 1$ to 2 **do**
  $v[n + 1, j] \leftarrow 0$
end for
for $i = 1$ to $n + 1$ **do**
  if $v[i, 1] > v[i, 2]$ **then**
    $tmp \leftarrow v[i, 1]$
    $v[i, 1] \leftarrow v[i, 2]$
    $v[i, 2] \leftarrow tmp$
  end if
end for
$div \leftarrow 2 \times \textsc{area}(v[\,,\,])$
return div

Algorithm 6 The empirical quantile comparison effect size, $\textsc{qces}(x[\,], y[\,])$

Require: a vector of observations $x[\,]$.
Require: a vector of observations $y[\,]$.
$div_{xy} \leftarrow \textsc{div}(x[\,], y[\,])$
$div_{yx} \leftarrow \textsc{div}(y[\,], x[\,])$
$qces \leftarrow \frac{1}{2} \times div_{xy} + \frac{1}{2} \times div_{yx}$
return $qces$